METHOD AND DEVICE FOR DISCRIMINATING VOICED AND UNVOICED SOUNDS

BACKGROUND OF THE INVENTION

1. Field of the Invention

This invention relates to a method and a device for making discrimination between the voiced sound and the noise or the unvoiced sound in speech signals.

2. Statement of Related Art

The speech or voice is classified into the voiced sound and the unvoiced sound. The voiced sound is the voice accompanied by vibrations of the vocal cord and consists in periodic vibrations. The unvoiced sound is the voice not accompanied by vibrations of the vocal cord and consists in non-periodic vibrations. The usual speech is composed mainly of the voiced sound, with the unvoiced sound being a special consonant termed unvoiced consonant. The period of the voiced sound is determined by the period of the vibrations of the vocal cord and is termed the pitch period, a reciprocal of which is termed a pitch frequency. In the following description, the term pitch means a pitch period. The pitch period and the pitch frequency are crucial factors on which depend highness or lowness of the speech or the intonation. Thus the sound quality of the speech depends on how precisely the pitch is grasped. However, in grasping the pitch, it is necessary to take account of the noise around the speech, or so-called background noise as well as quantization noise produced on quantization of analog signals into digital signals. In encoding speech signals, it is crucial to make distinction between the voiced sound from these noises and the unvoiced sound.

Among analog speech analysis systems, hitherto known in the art, there are such systems as disclosed in U.S. Pat. Nos. 4,637,046 and 4,625,327. In the former, input analog speech signals are divided into segments in the chronological sequence, and signals contained in these segments are rectified to find a mean value which is compared to a threshold value to make a voice/unvoiced decision. In the latter, analog speech signals are converted into digital signals and divided into segment and discrete Fourier transform is carried out from segment to segment to find an absolute value for each spectrum which is then compared to a threshold value to make a voiced/unvoiced decision.

Specific examples of encoding of speech signals include multi-band excitation coding (MBE), single band excitation coding (SBE), harmonic coding, sub-band coding (SBC), linear predictive coding (LPC), discrete cosine transform (DCT), modified DCT (MDCT) and fast Fourier transform (FFT).

For extracting the pitch from the input speech signal waveform by MBE coding, for example, pitch extraction may be achieved easily even if the pitch is not represented manifestly. For decoding at the synthesis side, a voiced sound waveform on the time domain is synthesized based on the pitch so as to be added to a separately synthesized unvoiced sound waveform on the time domain.

Meanwhile, if the pitch is adapted to be extracted easily, it may occur that a pitch that is not a true pitch be extracted in background noise segments. If such pitch other than the true pitch be extracted by MBE encoding, cosine waveform synthesis is performed so that peak points of the cosine waves are overlapped with one another at a pitch which is not the true pitch. That is, the cosine waves are synthesized by addition at a fixed phase (0-phase or $\pi/2$ phase) in such a manner that the voiced sound is synthesized at a pitch period which is not the true pitch period, such that the background noise devoid of the pitch is synthesized as a periodic impulse wave. In other words, amplitude intensities of the background noise, which intrinsically should be scattered on the time axis, are concentrated in a frame portion, with certain periodicity to produce an extremely obtrusive extraneous sound.

SUMMARY OF THE INVENTION

In view of the above-depicted status of the art, it is an object of the present invention to provide a method for making discrimination between voiced and unvoiced sounds whereby the voiced sound may positively be distinguished from the noise or unvoiced sound for preventing obtrusive extraneous sound from being produced during speech synthesis.

In one aspect, the present invention provides a method for discriminating a voiced sound from unvoiced sound or noise in input speech signals by dividing the input speech signals into blocks and giving a decision for each of these blocks as to whether or not the speech signals are voiced comprising the steps of subdividing one-block signals into a plurality of sub-blocks, finding statistical characteristics of the signals from one sub-block to another, and deciding whether or not the speech signals are voiced depending on a bias of the statistical characteristics on the time scale.

The peak value, effective value or the standard deviation of the signals for each of the sub-blocks may be employed as the aforementioned statistical characteristics.

In another aspect, the present invention provides a method for discriminating a voiced sound from an unvoiced sound or noise in input speech signals by dividing the input speech signals into blocks and giving a decision for each of these blocks as to whether or not the speech signals are voiced comprising the steps of finding the energy distribution of one-block signals on the frequency scale, finding the signal level of said one-block signals, and deciding whether or not the speech signals are voiced depending on the energy distribution and the signal level of one-block signals on the frequency scale.

Such voiced/unvoiced decision may also be made depending on the statistical characteristics of sub-block signals, namely the effective value, the standard deviation or the peak value and energy distribution of one block signals on the frequency scale, or alternatively, on the statistical characteristics of the sub-block signals, namely the effective value, the standard deviation or the peak value and the signal level of one-block signals.

In still another aspect, the present invention provides a method for discriminating a voiced sound from unvoiced sound or noise in input speech signals by dividing the input speech signals into blocks and giving a decision for each of these blocks as to whether or not the speech signals are voiced comprising the steps of subdividing one-block signals into a plurality of sub-blocks, finding statistical characteristics of the signals, that is effective value, standard deviation or peak value, from one sub-block to another, finding the energy distribution of the one-block signals on the frequency scale, finding the signal level of the one-block signals on the frequency scale, and deciding whether or not the speech signals are voiced depending on the effective value, standard deviation or the peak value, the energy distribution of the one-block signals on the frequency scale, and the signal level of the one-block signals on the frequency scale.